A Critical Appraisal of Techniques, Software Packages, and Standards for Quantitative Proteomic Analysis

Faviel F. Gonzalez-Galarza,¹ Craig Lawless,² Simon J. Hubbard,² Jun Fan,³
Conrad Bessant,³ Henning Hermjakob,⁴ and Andrew R. Jones¹

Abstract

New methods for performing quantitative proteome analyses based on differential labeling protocols or label-free techniques are reported in the literature on an almost monthly basis. In parallel, a correspondingly vast number of software tools for the analysis of quantitative proteomics data has also been described in the literature and produced by private companies. In this article we focus on the review of some of the most popular techniques in the field and present a critical appraisal of several software packages available to process and analyze the data produced. We also describe the importance of community standards to support the wide range of software, which may assist researchers in the analysis of data using different platforms and protocols. It is intended that this review will serve bench scientists both as a useful reference and a guide to the selection and use of different pipelines to perform quantitative proteomics data analysis. We have produced a web-based tool (http://www.proteosuite.org/?q=other_resources) to help researchers find appropriate software for their local instrumentation, available file formats, and quantitative methodology.

Introduction

IN CONTEMPORARY BIOLOGY, the development of technologies and methods to identify and quantify proteins has become an important issue for the understanding of biological systems (Weston and Hood, 2004). Identification and quantitation of proteins may assist in different analyses, for example for examining dissimilarities between samples or providing insights into the identification of potential biomarkers (Silberring and Caborowski, 2010). The use of mass spectrometry (MS) has been adopted as the gold standard for protein identification and now provides the basis for absolute and relative quantitation of peptides and proteins (Aebersold and Mann, 2003; Domon and Aebersold, 2006).

Two fundamental methodologies have prevailed in protein identification: (1) peptide mass fingerprint (PMF, MS¹ or simply MS), and (2) tandem mass spectrometry (MS² or MS/MS). PMF was the initial method for protein identification, focusing on the analysis of peptide masses; however, this approach is sub-optimal for the study of complex mixtures of proteins (Henzel et al., 2003; Thiede et al., 2005). To alleviate this problem, PMF was superseded by MS/MS techniques that are able to infer sequence information from spectra acquired from peptide fragmentation (see review in Domon and

Aebersold, 2006). These two methods depend on peptide separation normally performed by liquid chromatography (LC) in conjunction with MS, also called LC-MS (see review in Chen and Pramanik, 2008). Protein separation is often carried out prior to this step, by methods such as one dimensional (1D) or two dimensional (2D) gel electrophoresis, column chromatography, or affinity purification, followed by enzymatic digestion (often using trypsin) to generate the peptides.

Mass spectrometer output results, consisting of *m/z* and intensity values, are available in a wide variety of formats depending on the instrument vendors (e.g. '.raw' [Thermo Scientific], '.RAW' [Waters Corporation], and '.wiff' [Applied Biosystems – ABI], among many others). Raw files, which are usually binary encoded, are subsequently processed by different software packages to identify peptides and proteins covering different methodologies including (1) sequence database search [e.g., Mascot, OLAV (Colinge et al., 2003), OMSSA (Geer et al., 2004), ProBID (Zhang et al., 2002), SEQUEST (Eng et al., 1994), and X!Tandem (Craig and Beavis, 2004)], (2) peptide *de novo* sequencing [e.g., PEAKS[®] and MS BLAST (Shevchenko et al., 2001)], (3) tag searching [e.g., InsPect (Tanner et al., 2005)], and (4) spectral library searches [e.g., SpectraST (Lam et al., 2007) and X!Hunter (Craig et al., 2006)]. Several software packages have also been developed

¹Institute of Integrative Biology, University of Liverpool, Liverpool, U.K.

²Faculty of Life Sciences, University of Manchester, Manchester, U.K.

³Bioinformatics Group, Cranfield Health, Cranfield University, Cranfield, U.K.

⁴EMBL-The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, U.K.

to validate peptide assignments to MS/MS data produced by database searches, for example, PeptideProphet™ (Keller et al., 2002) and ProteinProphet™ (Nesvizhskii et al., 2003). (For peptide and protein identification software packages see reviews in McHugh and Arthur, 2008; Nesvizhskii, 2010; Xu and Ma, 2006).

The most common approach in proteomic pipelines is the database-driven search, in which observed peptide spectra are compared against theoretical tryptic peptides generated *in silico* using different algorithms. In this method, matching results for each candidate spectrum are subsequently ordered by some score or statistical measure, depending on the software used. Several algorithms often follow this by providing statistical estimates of confidence for peptide spectrum matches (PSMs), which model the likelihood of obtaining a given match by chance, such as false discovery rate (FDR) calculations using decoy databases (Elias and Gygi, 2007). Following peptide and protein identification, the challenge is to provide information on protein abundance for correct interpretation.

The need to characterize changes in gene expression at the protein level, from differences in protein abundance, has led to the design of different quantitation methodologies (see review in Ong and Mann, 2005). Each time a novel experimental method is devised in quantitative proteomics, new data interpretation algorithms also need to be designed and implemented. Thus on many occasions data analysis is fragmented, as existing software may only cover one or two methods, leading bench scientists to seek compatible tools/

pipelines to continue the analysis (Fig. 1A). This is a common problem, as most of the quantitation software tools do not support all search engines or instrument vendor's formats. Therefore, in this article we review some of the most widely used methods for quantitative proteomics, and describe an inventory of software tools that may be useful for data analysis. Recent reviews have also summarized some of the most popular quantitative proteomics applications (Lau et al., 2007; Mueller et al., 2008). However, we consider this a critical time for an update, with many recent advances in software for quantitative proteomics. The list encompasses free and proprietary software packages that are briefly discussed. Finally, the current status of community standards, which is the key in pipeline communication, is also included in this article.

Quantitation Techniques

As mentioned above, experimental techniques for identifying large numbers of proteins by MS are well established. One of the difficulties in using mass spectra directly to infer protein abundance is that MS is not inherently quantitative for peptide analysis; different peptides ionize to different degrees based on their specific biophysical properties, and prediction of signal based on peptide sequence is difficult. Quantitative proteomics thus relies on a diverse number of techniques and algorithms to assist in this challenge, based on differential labeling techniques, internal standards, targeted approaches via selected reaction monitoring (SRM), or complex

FIG. 1. Software pipelines in quantitative proteomics. Here a pipeline for performing quantitative analysis is shown. File format conversion tools and protein/peptide identification software packages are shown in panel A. A summary of quantitative techniques along with software packages (free and commercial) reviewed in this article are shown in panel B.

bioinformatics analysis for label-free quantitation. These multiple-dimension approaches are shown in Figure 1B. In the following section we present an outline of some of the most common quantitative techniques described in the literature (see also review in Bantscheff et al., 2007).

Labeling techniques

In quantitative proteomics, metabolic labeling is normally performed by comparing unlabeled (light) peptide to a different sample that is enriched with heavy stable isotopes of amino acids (e.g., Lys or Arg), or elements (e.g., ^{15}N or ^{18}O). One of the most common methods in metabolic labeling is the use of stable isotope labeling by amino acids in cell culture (SILAC). SILAC is a MS^1 -based method performed *in vivo* requiring cell culture; therefore this technique cannot be used in clinical samples. In SILAC experiments, two samples are normally mixed prior to digestion; using MS^1 signals to calculate the protein abundance (Ong et al., 2002). The differences between the intensity of the MS^1 features separated by the mass shift of the label gives an estimate of the relative abundance of the protein.

Other metabolic labeling techniques in quantitative proteomics include the isotope-coded affinity tag (ICAT), and the isotope-coded protein label (ICPL). The ICAT method relies on the use of chemical reagents that contain three main components: a reactive group that defines an amino acid side change, an isotopically-coded linker, and a tag for the isolation of labeled proteins or peptides (Gygi et al., 1999). As in other metabolic labeling strategies, to perform quantification, the two samples are labeled with light and heavy isotopes, respectively. Peptides are subsequently analyzed by LC-MS, and the ratios of signal intensities of the two sample pairs are calculated to determine the relative level of the protein in the two samples. In contrast, the ICPL technique (Schmidt et al., 2005) is based on stable isotope tagging at the frequent free amino groups of isolated proteins, which can be applied to different protein samples, for instance extracts from tissues. One advantage of ICPL is that it can be applied for multiplexed analysis in a single experiment.

Alternatively to isotopic approaches, the isobaric tag for relative and absolute quantitation (iTRAQ) has emerged as an MS^2 -based method. An iTRAQ tagging experiment comprises n -isobaric ($n=4,8$) reagents, each consisting of three groups: reporter, balance, and reactive groups. The total peptide pool from different samples is tagged with the different reagents (the reactive groups bind to the N-terminus of all peptides). Subsequently, the peptide pools from n -samples can be mixed prior to MS/MS analysis. A given peptide in different samples will thus have the same mass after being tagged in MS^1 (since the reagents are isobaric). However, during peptide fragmentation prior to MS^2 , the balance and reporter reagent are broken off from the peptide, and importantly, the different reporters ionize and can be detected, for example producing individual peaks in the mass range of 114, 115, 116, and 117 Da in the 4-plex system. The ratios of the intensity of ions detected for 114:115:116:117 are used to infer the relative differences between the amounts of each peptide in four samples running concurrently (Wiese et al., 2007). Additionally, other strategies for relative quantitation include tandem mass tags (TMTs), which is a similar technology to iTRAQ, developed by a different manufacturer (Thompson et al., 2003).

Label-free approaches

Label-free methods have gained popularity due to the low cost (i.e., there are no associated labeling reagent costs), and potential simplicity of laboratory protocols. These methods can be divided into two different strategies: (1) by measuring MS signal intensity of the peptide precursor ions, or (2) by counting the number of identified peptides matched to a given protein (spectral counting).

Spectral counting (SC)

In spectral counting, the number of peptide spectra matching a protein is considered to be an indicator of the protein abundance (Liu et al., 2004; Venable et al., 2004). The aim of SC is thus to quantify protein abundance by counting the number of detected tryptic peptides and their corresponding mass spectra. However, one of the limitations of SC is that this approach may bias protein abundance if only the number of peptide matches is considered, as larger proteins will produce more tryptic peptides and therefore more MS/MS spectra. In this technique, normalization of peptides is needed to account for different protein sequence lengths (number of tryptic peptides) and the ionization efficiency of different peptides. Inheriting the benefits of label-free approaches, researchers have used this method due to its scalability, easy implementation, and low cost compared to the label-based techniques. Analytical strategies for estimating the accuracy of FDR have also been performed in this approach (Lundgren et al., 2010).

Absolute protein expression (APEX)

In APEX, protein abundance is estimated from the number of PSMs associated with one protein. The APEX abundance is corrected via the prior expectation of observed peptides, using machine learning to calculate weighting factors for each PSM and each protein (Lu et al., 2007).

Exponentially-modified protein abundance index (emPAI)

In label-free analysis it is often necessary to normalize values of protein abundance to obtain a best approximation of protein quantitation. The exponentially-modified protein abundance index (emPAI) is another method for label-free analysis that is based on an optimization derived from the protein abundance index (PAI). PAI consists of the number of observed peptides divided by the theoretical observable peptides per protein (Rappsilber et al., 2002). In the case of emPAI, the formula is optimized by an exponential form of the PAI value (Ishihama et al., 2005). The correction reduces the overestimation of protein abundance generally observed by pure spectral count measures. Due to the relatively simple encoding task, the algorithm can be easily implemented in search engines (e.g., it is present in output files from Mascot).

Ion intensity

This label-free method is based on the quantitation of peptides according to their intensities, which are obtained from LC-MS/MS experiments. The technique generally involves the detection of features, defined as regions in two-dimensional LC-MS space corresponding to individual peptides, and the alignment of parallel LC-MS runs in the

retention time dimension, since LC is not highly reproducible (Fischer et al., 2006; Lange et al., 2007), to ensure that identical regions are compared across different runs (see also review in Lange et al., 2008a). These steps can occur in either order depending on the approach taken by the software package. The quantification of features can be performed by simple area-under-the-curve calculation for each isotope, or by fitting a model to the data, for example based on the average amino acid composition of a peptide, and quantifying the model rather than the data (which includes experimental noise). Other approaches include the Isotope Wavelet, whereby a model is fitted to the entire isotope pattern (Hussong et al., 2007), as implemented in the OpenMS software. As in other approaches, a database search must be performed from peak lists derived from MS² spectra to identify peptides, and a protein inference stage must be performed. Replication is particularly important in label-free methods, since increased variability is introduced by analyzing samples in parallel.

Average

Other methods for label-free analysis include chromatographic separations along with high-mass resolution and mass accuracy of time of flight (TOF) instruments (Silva et al., 2005). This technique, called Hi3, attempts to estimate absolute protein abundance based on the average intensity of the three most abundant peptide ions matched to each protein, using an internal spiked control for comparison. As with other ion intensity methods, comparison between accurate mass and retention time (AMRT) of two samples is performed to calculate the relative abundance.

Selected reaction monitoring (SRM)

Selected reaction monitoring (SRM), also known as multiple reaction monitoring (MRM), is a highly sensitive targeted technology commonly performed on triple-quadrupole (QQQ) mass spectrometers, that allows reliable quantitation of low-abundance proteins in complex mixtures (Lange et al., 2008b). In this technique, surrogate peptides are chosen for a set of target proteins of interest, from which quantitative data are acquired. Peptides are selected within the experimental context to be frequently observed when the protein is present (often referred to as proteotypic). Subsequently, selected MS reactions are defined using *m/z* filters, termed transitions, for a predefined precursor ion, leading to one or more specific fragment ions, typically using dedicated software tools [reviewed in Cham et al., 2010]. These transitions are expected to be uniquely characteristic for this given peptide ion, providing the high sensitivity. Quantitation values are subsequently obtained from intensities or peak areas of the XICs and compared across samples (relative), or to an isotopically-labeled internal standard (absolute).

Protein surrogates

AQUA. The absolute quantification (AQUA) method is a technique able to provide absolute protein quantitation by using chemically-synthesized, isotopically-labeled peptides as surrogates for a target protein of interest. The synthetic peptide is added to the sample mixture at a known concentration, which is digested and analyzed by LC-MS/MS (Gerber et al., 2003). Absolute values are then calculated by comparing XIC peak ratios of the target and reference peptides. AQUA is routinely coupled with the SRM strategy to absolutely quan-

tify a specific set of proteins. This approach can be expensive for large numbers of proteins owing to synthesis costs, though it does not have any digestion issues with respect to the reference peptide, and can be used for quantitation of post-translational modifications (Kirkpatrick et al., 2005).

QconCAT. The QconCAT method is a more economical solution for absolute quantitation of large numbers of proteins, such as in a systems biology context. Similarly to AQUA, the approach relies on selecting characteristic surrogate "Q-peptides" for the target set of proteins; however, they are then concatenated and encoded into a synthetic "QconCAT" gene (Beynon et al., 2005; Pratt et al., 2006). Briefly, the QconCAT is expressed in *Escherichia coli* in the presence of labeled media, purified, added to the protein sample at a known concentration, and submitted to proteolytic digestion and subsequent LC-MS or LC-MS/MS analysis. As with the AQUA approach, quantitation is obtained by calculating peak ratios between the reference and analyte, and it is ideally suited for SRM (Brownridge et al., 2011). This technique has the capacity to quantify ~50 peptides in a single sample; however, it can suffer from expression and digestion issues, so it is not particularly well suited for quantitation of post-translational modifications (PTMs). Both the QconCAT and AQUA approaches rely on the selection of proteotypic surrogate peptides for quantitation (Eyers et al., 2011); this is not always possible for each and every protein in a genome of interest.

PSAQ. The protein standard absolute quantification (PSAQ) strategy takes that QconCAT approach further by sub-cloning the full length of the target protein and expressing it in labeled media (Brun et al., 2009). As with a QconCAT the expressed protein is purified, but it can be spiked into the sample before any fractionation steps. This ensures that the reference and analyte protein/peptides behave the same within the protein mixture right through to the MS stage. Using the whole protein as a labeled reference not only ensures that the proteolytic digestion is identical between the reference and analyte, but also provides the maximum number of peptides from which to use for absolute quantitation. Similarly to QconCAT, it cannot usually be used for quantitation of PTMs, as the reference protein is unmodified.

SISCAPA. Finally, stable isotope standards and capture by anti-peptide antibodies (SISCAPA) is another method used for absolute quantitation of complex digests, comprising an important research tool for development of diagnostic markers (Anderson et al., 2004). This technique, which is designed for selected proteins of known sequence, can be useful in high-throughput assays for candidate disease marker proteins in plasma and body fluids and tissues.

Software for Quantitative Proteomics

Different software packages have been implemented to support the wide variety of quantitation techniques mentioned above (Fig. 1B). In this section, we describe some of the most common freely distributed and proprietary tools used in the analysis of quantitative proteomics data. A summary of software requirements, instruments, quantitation techniques and input file formats supported are shown in Table 1 (free software), and Table 2 (commercial software).

TABLE 1. FREE SOFTWARE PACKAGES FOR THE ANALYSIS OF QUANTITATIVE PROTEOMICS DATA

Software	Version	Date (month, year)	Technique	Type of data	Instruments	Input files	Distribution	License	Language	OS	URL	Reference
APEX	1.1.0	March10	Improvement of SC	LC-MS/MS	Any via proXML files	.fasta, .oi, proXML	Free open source	GNU GPL 3.0	Java	Windows, OSX, Linux	http://pfigr.cjvi.org/index.php/bioinformatics/apex.html	(Braisted et al., 2008)
Census	1.72	March 10	¹⁵ N, SILAC, iTRAQ, SC	MS ¹ /MS ²	Any via mzXML	MS1/MS2, DTASelect, mzXML, pepXML	Free	-	Java	Windows, OSX, Linux	http://fields.scripps.edu/census/	(Park et al., 2008)
MapQuant	2.1.1	Jan 06	Label free image recognition	LC-MS	Thermo, Waters	OpenRaw	Free open source	-	Visual C++	Windows, Linux	http://arep.med.harvard.edu/MapQuant/	(Leptos et al., 2006)
MaxIC-Q	0.9.2	-	SILAC, ICAT	LC-MS	Any via mzXML	mzXML, Mascot/SEQUEST/pepXML	Free	-	Visual C++	Windows, Web	http://ms64.iis.sinica.edu.tw/MaxIC-Q_web/index.html	(Tsou et al., 2009)
MaxQuant	1.2.2.5	-	SILAC, ICPL, Label free	MS ¹ /MS ²	LITQ, Orbitrap, FT-ICR (Thermo)	.raw (Thermo)	Free	-	C#	Windows	http://maxquant.org/	(Cox and Mann, 2008)
MFPaQ	4.0.0	-	SILAC, ICAT, label-free (SC and MS signals)	MS ¹ /MS ²	QSTAR XL, QSTAR Elite (ABI)	.wiff (ABI)	Free open source	-	Perl	Windows	http://mfpq.sourceforge.net/	(Bouysse et al., 2007)
mProphet	1.0.4.1	Dec 10	SRM, AQUA, QconCAT, PS-AQ	LC-MS/MS	Any via mzXML	mzXML + xls	Free open source	Apache 2.0	Perl/R	Windows, Linux	http://www.mprophet.org/	(Reiter et al., 2011)
MSQuant	2.0b6	May 10	¹⁵ N, SILAC, ICAT, ¹⁸ O, label-free	MS ¹ /MS ²	QSTAR (ABI), Q-ToF (Waters), LTQ, FT, Orbitrap (Thermo)	.raw (Thermo), .dat (Waters), .wiff (ABI)	Free open source	-	C# and VB .NET	Windows	http://msquant.alwaysdata.net/	(Mortensen et al., 2010)
Multi-Q	1.6.5.4	July 10	iTRAQ, SILAC, label-free	LC-MS/MS	Any via mzXML	mzXML, .wiff (ABI)	Free	-	VB .NET	Windows, Web	http://msiis.sinica.edu.tw/Multi-Q/	(Lin et al., 2006)
OpenMS	1.8	March 11	iTRAQ, SILAC, label-free	MS ¹ /MS ²	Any via mzXML or mzML	.data, mzData, mzXML, mzML	Free open source	LGPL	C++	Windows, Linux, OSX	http://open.ms.sourceforge.net/	(Kohlbaicher et al., 2007)
PeakQuant	1.5.42	-	¹⁵ N, SILAC, iTRAQ	MS ¹ /MS ²	Any via mzXML	mzXML	Free	-	Java	Windows, Linux, OSX	http://www.medicinisches-proteom-center.de/software	Included in this issue
PEPPER	-	-	Label-free	MS ¹ /MS ²	Any via mzXML	mzXML	Free	-	Perl	Windows	http://www.broadinstitute.org/cancer/software/genepattern/desc/proteomics	(Jaffe et al., 2006)
ProRata	1.0	Dec 05	¹⁵ N, SILAC, ICAT, ¹⁸ O	LC-MS	Any via mzXML	mzXML	Free open source	GNU GPL 3.0	C++	Windows, Linux	http://code.google.com/p/prorata/	(Pan et al., 2006)
Proteios	2.16	June 11	iTRAQ, TMT	MS ¹ /MS ²	Any via mzML	mzML	Free open source	GPL v2	Java	Windows, Linux, OSX	http://www.proteios.org/	(Halkinen et al., 2009)
QUIL	-	-	¹⁸ O, ICAT	LC-MS	LCQ, LTQ, FT-ICR (Thermo)	-	Available on request	-	Visual C++	Windows	-	(Wang et al., 2006)
Qupe	-	-	¹⁵ N	LC-MS	LTQ, Orbitrap (Thermo)	mzXML	Web	-	Java	Web-based	https://qupe.cebiitec.uni-bielefeld.de/	(Albaum et al., 2009)
Skyline	1.2	March12	SRM, label-free	LC-MS	Any via mzXML	mzXML, pepXML	Free open source	Apache 2.0	C#	Windows	http://proteome.gs.washington.edu/software/skyline	(MacLean et al., 2010)
STEM	-	-	¹⁸ O	LC-MS	-	.pkl (ProteinLynx, Waters)	-	-	-	Windows	-	(Shinkawa et al., 2005)
TPP	4.5.0	Aug 11	ICAT, SILAC, iTRAQ	MS ¹ /MS ²	See below	mzXML, mzML	Free	-	-	Windows, Linux, OSX	http://www.proteomecenter.org/software.php	(Deutsch et al., 2010)
- ASAPRatio	-	-	ICPL, ICAT, SILAC	-	Any via mzXML or mzML	Via TPP	Free	-	C	Windows, Linux, OSX	http://tools.proteomecenter.org/wiki/	(Li et al., 2003)
- Libra	-	-	iTRAQ	-	Any via mzXML or mzML	Via TPP	Free	-	C	Windows, Linux, OSX	http://tools.proteomecenter.org/wiki/	(Pedrioli et al., 2004)
- SpecArray	-	-	Label-free	-	Any via mzXML or mzML	Via TPP	Free	-	C	Linux	http://tools.proteomecenter.org/wiki/	(Li et al., 2005)
- SuperHirn	-	-	Label-free	-	Any via mzXML or mzML	Via TPP	Free	-	C++	Linux, OSX	http://tools.proteomecenter.org/wiki/	(Mueller et al., 2007)
- XPRESS	-	-	ICPL, ICAT, SILAC, ¹⁴ N/ ¹⁵ N	-	Any via mzXML or mzML	Via TPP	Free	-	C	Windows, Linux, OSX	http://tools.proteomecenter.org/wiki/	(Han et al., 2001)
VIPER	3.48.456	Sept 11	¹⁸ O, ICAT	MS ¹ /MS ²	Any via mzXML	.pek, CSV, mzXML, mzData, .raw (Thermo)	Free open source	Apache 2.0	-	Windows	http://omics.pnl.gov/software/VIPER.php	(Monroe et al., 2007)
X-Tracker	1.3	June 11	iTRAQ, ¹⁵ N	MS ¹ /MS ²	Any via mzXML	mzML, mzIdentML	Free open source	GNU GPL 3.0	Java	Windows, Linux, OSX	http://www.x-tracker.info/	-
- iTracker	1.1	May 05	iTRAQ	MS ²	-	.mgf, .dta	Free	-	Perl	Windows, Linux	http://www.cranfield.ac.uk/health/research/charas/bioinformatics/page6801.html	(Shadforth et al., 2005)
ZoomQuant	-	-	¹⁸ O	LC-MS	LTQ (Thermo)	.raw (Thermo)	Free	-	Perl	Windows, Linux, OSX	http://proteomics.mcu.edu/zoomquant.html	(Halligan et al., 2005)

List as of April 30, 2012.

SC, spectral count; (-), information not available; LGPL, lesser GNU Public License; LC, liquid chromatography MS, mass spectrometry; OS, operating system; URL, uniform resource locator.

TABLE 2. COMMERCIAL SOFTWARE PACKAGES FOR THE ANALYSIS OF QUANTITATIVE PROTEOMICS DATA

Software	Company	SILAC	iTRAQ	¹⁵ N	¹⁸ O	ICPL	ICAT	LF	TMT	Other
BioWorks	Thermo	Y	Y				Y	Y		
Elucidator	Rosetta	Y					Y	Y		
Mascot Distiller	Matrix Science	Y			Y	Y	Y	Y		Absolute quantitation (AQUA)
PEAKS Q	Bioinformatics Solutions	Y	Y		Y	Y	Y	Y	Y	
PLGS	Waters							Y		Absolute quantitation (Hi3)
Progenesis LC-MS	NonLinear Dynamics							Y		
Pro Quant	Applied Biosystems		Y							
ProteinPilot	AB SCIEX	Y	Y			Y	Y			
ProteoIQ	Bioinquire	Y	Y	Y	Y		Y	Y	Y	
Proteome Discoverer	Thermo		Y						Y	Absolute quantitation (HeavyPeptide - AQUA)
SIEVE	Thermo							Y		
WARP-LC	Bruker	Y				Y				

List as of April 30, 2012.

LF, label-free based on precursor ion intensities.

Free distributions

APEX is an open-source application written in Java based on the Absolute Protein Expression algorithm (Braisted et al., 2008). Containing a graphical user interface, APEX computes protein abundance from LC-MS/MS data. The latest version (1.1.0, available since March 2010) performs label-free quantitation via the normalization of spectral count data. APEX takes as input protXML files produced by the ProteinProphet module included in the TPP pipeline (Deutsch et al., 2010; Keller et al., 2005). In this technique, machine learning algorithms are used to predict weighting factors for each PSM based on the predicted biophysical properties of the peptide. Then the spectral count is weighted accordingly and used to calculate the protein abundance.

Census is a free Java application that has the flexibility to perform labeling and label-free quantitation analysis (Park et al., 2008). As of version 1.72, Census supports ¹⁵N, SILAC, iTRAQ, and spectral counting for label-free through the use of chromatogram alignments. To eliminate the noise in peak detection, Census uses two different approaches depending on data resolution: by examining the whole isotope envelope (low resolution), or analyzing individual isotopes (high resolution). Census requires either MS1/MS2 file formats (McDonald et al., 2004) or mzXML as input. MS1/MS2 files can be produced via the RawXtractor application, which supports Thermo Fisher instruments only. In contrast, mzXML files can be generated for various instruments using different converters (e.g., CompassXport-Bruker, ReAdW-Thermo, Wolf-Waters, and mzWiff-ABI).

MapQuant is an open-source software package implemented in Visual C++ that is used to quantify proteins in large datasets by image recognition (Leptos et al., 2006). In this software, techniques of peak detection (peak finding, fitting, and clustering), and noise filtering in the MS¹ dimension are also performed. To be instrument-independent, raw files must be converted into OpenRaw format. To do this, a converter from XCalibur (Thermo Scientific) to OpenRaw is provided in the MapQuant website.

MaXIC-Q provides quantitation analysis of large-scale datasets supporting SILAC, ICAT, and other labeling methods produced by users (Tsou et al., 2009). Implemented in Visual C++,

the latest release (0.9.2) can be run in two modes, via web or as a desktop application. The algorithm implemented in MaXIC-Q is based on areas of XICs to determine ion abundance, which is used in the estimation of protein and peptide ratios.

MaxQuant software, produced by Matthias Mann's group, has become a well-established program for the analysis of quantitative data performing SILAC (Cox and Mann, 2008; Cox et al., 2009). In addition to SILAC, the last release (1.2.2.5) of MaxQuant supports ICPL and label-free methods. In MaxQuant, peaks are detected by fitting the peak shape to three central data points, which are assembled into three dimensional peak hills over the *m/z* axis. For MaxQuant, input data is provided in raw files generated by Thermo Finnigan mass spectrometers. One limitation of MaxQuant is that it can only be run in Windows.

MFPaQ is a free and open-source program written in Perl that supports SILAC, ICAT, and label-free, by spectral count and signal comparisons (Bouyssié et al., 2007). To run this software, the popular Mascot search engine is needed. This software platform also allows the organization of Mascot identifications by using customized filters from the user. Additionally, MFPaQ permits the calculation of peptide ratios, and generates data for non-redundant proteins identified, along with the average and normalized ratios. One of the drawbacks of this software is that an Internet Information Services (IIS) web server is required, restricting the usability to Windows users only.

mProphet is an open-source suite developed in Perl and R, specifically designed for automated processing of large scale SRM/MRM experiments (Reiter et al., 2011). The program takes mzXML files as input and combines all of the SRM measurements into a probabilistic scoring model. This scoring model is used to assess the quality of peaks in the XIC for the target peptide; including co-elution, peak shape, and relative intensities of fragment ions. Peak intensities and AUCs are calculated for all target (and reference) peptides, which are subsequently used to calculate peak ratios for relative or absolute quantitation, dependent on the aim of the experiment. In addition, mProphet requires the acquisition of decoy transition data within the experiment. These decoy data are then used as negative controls to provide FDR scores to assess the quality of the real data. The software was developed using

Thermo and ABI machines, but can readily be used on any raw SRM data that can be converted into mzXML format. The mProphet package is able to run on both Linux and Windows machines, or alternatively can be obtained pre-installed on a virtual machine.

MSQuant is an open-source environment that supports relative quantitation based on precursor ion intensities (Mortensen et al., 2010). The software supports ^{15}N , SILAC, ICAT, ^{18}O , iTRAQ, and label-free ion intensity methods. This application was intended to allow recalibration by increasing mass accuracy and reducing false-positive peptide identifications. This software was written in C# and VB.NET. Distributed as vendor-independent, MSQuant supports .raw, .dat, and .wiff files from Thermo, Waters, and Applied Biosystems; however, it is also restricted to Windows machines only.

Multi-Q is a software package developed in VB.NET for multiplexed quantitation of protein expression supporting the iTRAQ technique (Lin et al., 2006; Yu et al., 2007). To calculate ratios, Multi-Q performs filtering of peptides that match different proteins. One advantage of the software is that it contains some statistical methods to handle low-resolution MS/MS spectra. The software also provides a data converter for processing spectral raw data generated by the majority of mass spectrometer manufacturers (ABI, Thermo, Waters, and Bruker). Additionally, the package supports results from Mascot, SEQUEST, and X!Tandem.

OpenMS is another popular open-source suite designed for identification, raw data processing tasks, and quantitation analysis (Kohlbacher et al., 2007). The Open MS Proteomics Pipeline (TOPP) consists of several applications that can be arranged to create an analysis pipeline to target a specific problem. Several implementations have been included in this software package to perform quantitative analysis for a particular technique, such as the SILACAnalyser (Nilse et al., 2010), and iTRAQAnalyser, using the mzML data standard (see section below on data standards) as a common input to all modules. This software package includes different readers that support the majority of types of raw files. The OpenMS project is developed in C++, and can be compiled in Windows, Linux, and OSX machines.

PeakQuant is another free Java-based application for the analysis of quantitative data. This integrated platform allows quantitative analysis of different experiments, including $^{14}\text{N}/^{15}\text{N}$, SILAC, and iTRAQ. One of the modules integrated into this package is a generic algorithm called FindPairs, which is considered to provide high accuracy and reliability of data by using isotope-coded mass spectra. PeakQuant can also read mzXML, providing users with a non-dependent instrument pipeline support.

PEPPER is a platform for experimental proteomics pattern recognition (Jaffe et al., 2006). This software uses two main algorithms, landmark matching and peak matching, in order to reduce the problems of reproducibility in LC. Landmark matching is a method to propagate identified peptides over time onto accurate mass LC-MS features, allowing the trace of the total identified peptides from disparate data acquisition methods. Peak matching is then used to recognize identical molecular species across multiple LC-MS experiments by clustering. PEPPER has been designed for optimal performance on high-resolution data, but may be adaptable to different instrument types. This software package is written in Perl and distributed for Windows machines only.

ProRata is a computer program that automates data processing for quantitative shotgun proteomics (Pan et al., 2006). SILAC, ^{15}N , ICAT, and ^{18}O techniques are supported in this software. ProRata takes the output from SEQUEST via the DTASelect program along with the mass spectral data and performs all steps necessary for estimating protein abundance ratios. To perform the analysis, ProRata extracts data for selected ion chromatograms, detects chromatographic peaks, and estimates peptide and protein abundance ratios. As the source code is implemented in C++, ProRata can be compiled in Linux and Windows.

Proteios is a web-based local data repository for proteomics experiments (Hakkinen et al., 2009). The package contains plug-ins that allow iTRAQ and TMT quantitation. The current stable version (2.6) is one of the few standard-compliant software packages supporting mzML files. Proteios is distributed as an open-source Java code supporting different platforms.

QUIL is a program for automated protein quantitation based on ion chromatogram reconstruction (Wang et al., 2006). ICAT and ^{18}O techniques have been tested in several instruments, such as LCQ, LTQ, and FT-ICR. QUIL also provides a graphical interface to visualize chromatograms and facilitate the validation of peak detection and ratio calculation.

Qupe is a web-based application written in Java that allows data management and analysis functions for LC-MS/MS experiments (Albaum et al., 2009). The application allows importing mass spectra data and performs peptide identification using a database search. The Qupe environment, supporting ^{15}N , is provided as a web service to avoid desktop installations.

Skyline is a free open-source software package for quantitative proteomics data analysis mainly designed for targeted proteomics (MacLean et al., 2010). This program was initially distributed with support of SRM; however, recent releases also include the option for label-free analysis (Schilling et al., 2012). Skyline is developed in C#, making the software a Windows-based application.

The Strategic Extractor for Mascot's results (STEM) software is a stand-alone tool for the analysis of quantitative data (Shinkawa et al., 2005). STEM supports ^{18}O labeling, and is mainly designed for Windows machines. To perform the analysis, this software requires .pkl (Micromass ProteinLynx, Waters) and Mascot files as inputs.

The Trans-Proteomics Pipeline (TPP) is a software suite distributed by the Seattle Proteome Centre that performs identification and quantitation analysis (Deutsch et al., 2010). Different options are provided within the TPP to perform a particular quantitation analysis. For example, ASAPRatio is used for converting peptide level information to protein-level ratios, and supports ICPL, ICAT, and SILAC techniques (Li et al., 2003), SuperHirn, for label-free (Mueller et al., 2007), SpecArray, also for label-free by the use of peptide arrays (Li et al., 2005), Libra, designed for iTRAQ technique (Keller et al., 2005), and XPRESS, used for ^{15}N , ICPL, ICAT, and SILAC labeling data (Han et al., 2001). The TPP requires the installation of several software packages such as Apache to run in a web-based environment.

VIPER is a free open-source graphical software that supports the ^{18}O and ICAT quantitation methods (Monroe et al., 2007). This software supports high-throughput peptide identification by accurate mass and time tag approach, and can run on Windows-based machines requiring Microsoft Access to create AMT tag databases. As input files, the

software can read .raw (Thermo) files, and the previous standard formats mzXML and mzData.

The i-Tracker program was designed as a counterpart of the Pro Quant Software provided by Applied Biosciences (Shadforth et al., 2005). This software has been developed at Cranfield University to extract the quantitative information in a format easily linked to other protein identification tools such as Mascot and SEQUEST. In order to support additional techniques, a new implementation called X-Tracker was developed by the Cranfield group. The aim of X-Tracker is to provide a plug-in-based framework to support some of the most common quantitation methods.

ZoomQuant is another quantitative software package that supports ^{18}O technique (Halligan et al., 2005). This standalone application, written in Perl, analyzes the mass spectra of ^{18}O -labeled peptides from ion trap instruments, and determines the relative abundance between the two samples. ZoomQuant can read .raw files from Thermo Instruments and run in multi-platforms.

Commercial software

Alternatively to open-source and/or freely available tools, commercial software provides excellent tools that are focused on particular quantitation techniques, instruments, formats, and software dependencies (Table 2). BioWorks™ (Thermo Scientific) supports SILAC, iTRAQ, ICAT, and label-free quantitation techniques. The identification of proteins is based on the SEQUEST searching algorithm by comparing MS/MS spectra data against protein and DNA databases. Elucidator® (Rosetta) supports label-free and SILAC and ICAT metabolic labeling for different instrument vendors such as Waters, Thermo, Applied Biosystems, and Agilent Technologies. Mascot Distiller (Matrix Science) detects peaks to fit an ideal isotopic distribution to the experimental data, which is predicted from the elemental composition for a peptide of average amino acid composition. At present, it supports ^{18}O , AQUA, ICAT, ICPL, metabolic labeling SILAC, and label-free, based on the quantitation of relative intensities which are extracted from ion chromatograms. PEAKS® Q (Bioinformatics Solutions) is based on feature detection along with noise filtering. This software package calculates the relative abundance of proteins, supporting ICAT, ICPL, iTRAQ, and label-free quantitation, based on the use of intensities of precursor ions, N-terminal labeling, ^{18}O , SILAC, TMT, or other user-defined labeling. ProteinLynx Global SERVER™ (Waters) is a suite based on the use of exact mass measured data (using absolute quantitation), and specificity of data-independent acquisition (MS^E) analysis. Progenesis LC-MS™ (NonLinear Dynamics) is mainly designed for label-free analysis, through quantitation of ion abundance. The current version can support most of the major machine vendors and databases. Pro Quant (Applied Biosystems) enables quantitation and identification of iTRAQ-labeled peptides. Peptide identification is achieved using the Interrogator™ search algorithm from ABI. ProteinPilot™ (AB SCIEX) works for iTRAQ, ICPL, ICAT, and SILAC. ProteoIQ (Bioinquire) provides label-free quantitation by spectral counting using TIC, XIC, or spectral counting. It also supports isobaric tag quantitation (iTRAQ and TMT), and isotopic labeling quantitation (SILAC, ICAT, ^{18}O , ^{15}N , dimethyl labeling, and acetylation). Proteome Discoverer (Thermo Scientific) supports isobaric

mass tagging (TMT and iTRAQ), and relative and absolute quantitation, by using HeavyPeptide techniques. SIEVE (Thermo) is the software solution for label-free analysis. Using MS intensities from raw LC-MS data, this software employs an algorithm called ChromAlign to reduce the effect of chromatographic variability between samples. WARP-LC (Bruker) supports isobaric and isotopic labeling, including ICPL and SILAC.

Selecting a Software Pipeline

As reviewed in the previous section, the process of selecting the optimal software solution is not a simple task, and pipelines may be restricted to instrument, protocol, or software/hardware availability. While commercial packages may provide efficient algorithms for the analysis of quantitative data, most of the software is restricted to a particular instrument or specific quantitation strategy. To aid bench scientists in the selection of existing software pipelines, we have produced a user-configuration pipeline tool (http://www.proteosuite.org/?q=other_resources; Fig. 2). As shown in the example, users can select a particular instrument, standard file format, or technique, and a list will be displayed with the different software packages that fit the criteria.

Community Standards in Proteomics

One of the drawbacks in the past was the lack of a common format to represent data produced from instrument vendors containing mass spectra data. Usually, files were binary encoded, and it was difficult for different software packages to cover all the different instruments. Two developments have started to improve capabilities for open analysis. First, the Institute for Systems Biology produced a set of open XML-based file formats within the TPP that have become popular for open-source development, including mzXML (raw or processed mass spectra), pepXML (peptide identifications and quantitative values), and protXML (protein identifications and quantitative values Keller et al., 2005; Pedrioli et al., 2004). Second, the development of community data standards has been led by the Human Proteome Organisation (HUPO) via the Proteome Standards Initiative (PSI Orchard et al., 2003). Several standard formats have been developed and provide support for the different protein identification and quantitation steps. Although this is still in development, some mature definitions have been agreed upon. In early work, the PSI developed the mzData format for mass spectra and related metadata. At that time, there was parallel development between mzData and the ISB's mzXML. The ISB and PSI agreed to work on a single mass spectral standard—called mzML. The current version 1.1 has been available since June 2009 (Martens et al., 2011). Since then, several packages have been incorporated this standard, for example, ProteoWizard (Kessner et al., 2008) which contains a set of libraries that can convert most of the different vendor instrument files into mzML. Some software packages have used these or their own libraries to support mzML files directly. Several libraries have been developed by different groups to convert their mass spectra into mzML format (e.g., a Java implementation for mzML files called jmzML; Cote et al., 2010). Other tools such as TPP, X!Tandem, Proteios, and Mascot, among many others, have adopted this standard in their packages (a full list can be found at <http://www.psidev.info/mzML>).

FIG. 2. Screen shot of the pipeline configurator tool on www.proteosuite.org. Shown here is an example of the list of software packages based on the type of instrument (file format). A list of quantitative techniques is also provided to refine the search.

The PSI has developed a separate format to describe peptide and protein identification data, along with metadata to describe how the identifications were made, in the mzIdentML format. The current version 1.1 has been available since August 2011 (Jones et al., 2012). The mzIdentML format has been adopted as a native standard format for several commercial software packages, including Mascot, Scaffold, and Phenyx, and open-source parsers or file converters have been developed for OMSSA, X!Tandem, and OpenMS, among others (a full list can be found at <http://www.psdev.info/tools-implementing-mzidentml>). Additionally, mzIdentML support is provided in ProteoWizard, and a Java API, jmzIdentML, has recently been released (Reisinger et al., 2012).

To handle and describe quantitative data produced in proteomics studies, the PSI has defined the mzQuantML data standard. The standard captures the output of quantitative software, defining values associated with features identified within two-dimensional LC-MS space, plus quantitative values associated with peptides, proteins, or protein groups, matched across different experimental conditions. The development of mzQuantML is ongoing, but a version 1.0 release is expected in 2012 (<http://code.google.com/p/mzquantml/>). Another format has been designed specifically for supporting SRM studies—the PSI TraML standard (Deutsch et al., 2012)—which has been developed to allow seamless exchange of transition lists between software tools, instruments, and databases (<http://psdev.info/trauml>).

Future Perspectives in Quantitative Proteomics Software

As reviewed in this article, a vast number of software packages and algorithms have been implemented over the last 10 years. However, few software packages can cover all the different quantitative techniques requiring intermediate applications (if they exist) to continue the flow when performing data analysis. Suffice it to say that the need to develop software pipelines is essential. To do this, community

standard formats have emerged as a solution to transparently interconnect different software packages. Additionally, by the use of standard formats, bench scientists can be provided with standard datasets that can be submitted to web repositories for re-analysis. Recent open-source APIs have been provided to assist individuals in the implementation of standard formats for a particular programming language (e.g., jmzml and jmzidentml for Java; Cote et al., 2010; Reisinger et al., 2012; or the ProteoWizard utilities implemented in C++; Kessner et al., 2008). By promoting software reusability, scientists may contribute to software optimization and tackle one of the biggest problems found in free distribution software packages, namely continuing support. In an effort to circumvent this problem, consortiums such as the ProteoSuite project (www.proteosuite.org) aim to provide continuing development and targeting of the coverage of most of the techniques described here. It is intended that the contribution of different specialist algorithms for each module (e.g., raw data reading, peptide/protein identification, peak selection, and noise filtering) will enhance quantitation methods, and consequently improve interpretation of results. It is also envisioned that future developments in quantitative proteomics will improve user-friendliness by providing a one-click installation process and suitable graphical interfaces, thus facilitating their use by non-expert users. Finally, in many cases, publishers of software packages do not disclose the algorithms used in their construction, making it difficult to evaluate the performance of the results. Thus, the development of well-documented open-source tools for proteomics data analysis would benefit scientists in the generation of more accurate results.

Conclusions

In this article we have provided a review of the most widely used methods in quantitative proteomics, and described some of the software packages available to perform the analysis of data. Benchmarking of the different software packages is always a demanding task, and can lead to bias in the selection of

particular software, as some of the packages are method-specific (Yates et al., 2012). Therefore, rather than selecting only one software solution, the aim of this review was to help scientists in the design of the different configuration pipelines according to resource availability, highlighting the use of community standards as a method to communicate the different modules.

Acknowledgments

This work was supported by the British Biotechnology and Biological Sciences Research Council (grants BB/I00095X/1, BB/I000909/1, BB/I000631/1, and BB/I001131/).

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

References

- Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* 422, 198–207.
- Albaum, S.P., Neuweiger, H., Franzel, B., et al. (2009). Qupe—a Rich Internet Application to take a step forward in the analysis of mass spectrometry-based quantitative proteomics experiments. *Bioinformatics* 25, 3128–3134.
- Anderson, N.L., Anderson, N.G., Haines, L.R., Hardie, D.B., Olafson, R.W., and Pearson, T.W. (2004). Mass spectrometric quantitation of peptides and proteins using stable isotope standards and capture by anti-peptide antibodies (SISCAPA). *J Proteome Res* 3, 235–244.
- Bantscheff, M., Schirle, M., Sweetman, G., Rick, and J., Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. *Analyt Bioanal Chem* 389, 1017–1031.
- Beynon, R.J., Doherty, M.K., Pratt, J.M., and Gaskell, S.J. (2005). Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat Meth* 2, 587–589.
- Bouyssié, D., De Peredo, A.G., Mouton, E., et al. (2007). Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses. *Mol Cell Proteomics* 6, 1621–1637.
- Braisted, J.C., Kuntumalla, S., Vogel, C., et al. (2008). The APEX Quantitative Proteomics Tool: Generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics* 9, 529.
- Brownridge, P., Holman, S.W., Gaskell, S.J., et al. (2011). Global absolute quantification of a proteome: Challenges in the deployment of a QconCAT strategy. *Proteomics* 11, 2957–2970.
- Brun, V., Masselon, C., Garin, J.R.M., and Dupuis, A. (2009). Isotope dilution strategies for absolute quantitative proteomics. *J Proteomics* 72, 740–749.
- Cham, J.A., Bianco, L., and Bessant, C. (2010). Free computational resources for designing selected reaction monitoring transitions. *Proteomics* 10, 1106–1126.
- Chen, G.D., and Pramanik, B.N. (2008). LC-MS for protein characterization: current capabilities and future trends. *Expert Rev Proteomic* 5, 435–444.
- Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003). OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics* 3, 1454–1463.
- Cote, R.G., Reisinger, F., and Martens, L. (2010). jmzML, an open-source Java API for mzML, the PSI standard for MS data. *Proteomics* 10, 1332–1335.
- Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26, 1367–1372.
- Cox, J., Matic, I., Hilger, M., et al. (2009). A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat Protoc* 4, 698–705.
- Craig, R., and Beavis, R.C. (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466–1467.
- Craig, R., Cortens, J.C., Fenyo, D., and Beavis, R.C. (2006). Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res* 5, 1843–1849.
- Deutsch, E.W., Chambers, M., Neumann, S., et al. (2012). TraML—a standard format for exchange of selected reaction monitoring transition lists. *Mol Cell Proteomics* 11, R111 015040.
- Deutsch, E.W., Mendoza, L., Shteynberg, D., et al. (2010). A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 10, 1150–1159.
- Domon, B., and Aebersold, R. (2006). Review—Mass spectrometry and protein analysis. *Science* 312, 212–217.
- Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 4, 207–214.
- Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J Am Soc Mass Spectr* 5, 976–989.
- Eyers, C.E., Lawless, C., Wedge, D.C., Lau, K.W., Gaskell, S.J., and Hubbard, S.J. (2011). CONSeQuence: Prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. *Mol Cell Proteomics* 10, M110.003384. Epub 2011 Aug 3.
- Fischer, B., Grossmann, J., Roth, V., Gruissem, W., Baginsky, S., and Buhmann, J.M. (2006). Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics* 22, E132–E140.
- Geer, L.Y., Markey, S.P., Kowalak, J.A., et al. (2004). Open mass spectrometry search algorithm. *J Proteome Res* 3, 958–964.
- Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W., and Gygi, S.P. (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci USA* 100, 6940–6945.
- Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 17, 994–999.
- Hakkinen, J., Vincic, G., Mansson, O., Warell, K., and Levander, F. (2009). The Proteios Software Environment: An Extensible multiuser platform for management and analysis of proteomics data. *J Proteome Res* 8, 3037–3043.
- Halligan, B.D., Slyper, R.Y., Twigger, S.N., Hicks, W., Olivier, M., and Greene, A.S. (2005). ZoomQuant: An application for the quantitation of stable isotope labeled peptides. *J Am Soc Mass Spectr* 16, 302–306.
- Han, D.K., Eng, J., Zhou, H.L., and Aebersold, R. (2001). Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat Biotechnol* 19, 946–951.
- Henzel, W.J., Watanabe, C., and Stults, J.T. (2003). Protein identification: the origins of peptide mass fingerprinting. *J Am Soc Mass Spectrom* 14, 931–942.
- Husson, R., Tholey, A., and Hildebrandt, A. (2007). Efficient analysis of mass spectrometry data using the isotope wavelet. *AIP Conf Proc* 940, 139–149.

- Ishihama, Y., Oda, Y., Tabata, T., et al. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* 4, 1265–1272.
- Jaffe, J.D., Mani, D.R., Leptos, K.C., Church, G.M., Gillette, M.A., and Carr, S.A. (2006). PEPpeR, a platform for experimental proteomic pattern recognition. *Mol Cell Proteomics* 5, 1927–1941.
- Jones, A.R., Eisenacher, M., Mayer, G., et al. (2012). The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics*.
- Keller, A., Eng, J., Zhang, N., Li, X.J., and Aebersold, R. (2005). A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 1:2005.0017. Epub 2005 Aug 2.
- Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analyt Chem* 74, 5383–5392.
- Kessner, D., Chambers, M., Burke, R., Agusand, D., and Mallick, P. (2008). ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24, 2534–2536.
- Kirkpatrick, D.S., Gerber, S.A., and Gygi, S.P. (2005). The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods* 35, 265–273.
- Kohlbacher, O., Reinert, K., Gropl, C., et al. (2007). TOPP—the OpenMS proteomics pipeline. *Bioinformatics* 23, e191–197.
- Lam, H., Deutsch, E.W., Eddes, J.S., et al. (2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7, 655–667.
- Lange, E., Gropl, C., Schulz-Trieglaff, O., Leinenbach, A., Huber, C., and Reinert, K. (2007). A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics* 23, i273–i281.
- Lange, E., Tautenhahn, R., Neumann, S., and Gropl, C. (2008a). Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* 9, 375.
- Lange, V., Picotti, P., Domon, B., and Aebersold, R. (2008b). Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* 4, 222.
- Lau, K.W., Jones, A.R., Swainston, N., Siepen, J.A., and Hubbard, S.J. (2007). Capture and analysis of quantitative proteomic data. *Proteomics* 7, 2787–2799.
- Leptos, K.C., Sarracino, D.A., Jaffe, J.D., Krastins, B., and Church, G.M. (2006). MapQuant: Open-source software for large-scale protein quantification. *Proteomics* 6, 1770–1782.
- Li, X.J., Yi, E.C., Kemp, C.J., Zhang, H., and Aebersold, R. (2005). A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol Cell Proteomics* 4, 1328–1340.
- Li, X.J., Zhang, H., Ranish, J.A., and Aebersold, R. (2003). Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Analyt Chem* 75, 6648–6657.
- Lin, W.T., Hung, W.N., Yian, Y.H., et al. (2006). Multi-Q: A fully automated tool for multiplexed protein quantitation. *J Proteome Res* 5, 2328–2338.
- Liu, H., Sadygov, R.G., and Yates, J.R. 3rd (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analyt Chem* 76, 4193–4201.
- Lundgren, D.H., Hwang, S.I., Wu, L.F., and Han, D.K. (2010). Role of spectral counting in quantitative proteomics. *Expert Rev Proteomic* 7, 39–53.
- Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E.M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 25, 117–124.
- Maclean, B., Tomazela, D.M., Shulman, N., et al. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26, 966–968.
- Martens, L., Chambers, M., Sturm, M., et al. (2011). mzML—a community standard for mass spectrometry data. *Mol Cell Proteomics* 10, R110 000133.
- McDonald, W.H., Tabb, D.L., Sadygov, R.G., et al. (2004). MS1, MS2, and SQT—three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun Mass Sp* 18, 2162–2168.
- McHugh, L., and Arthur, J.W. (2008). Computational methods for protein identification from mass spectrometry data. *PloS Comput Biol* 4, e12.
- Monroe, M.E., Tolic, N., Jaitly, N., Shaw, J.L., Adkins, J.N., and Smith, R.D. (2007). VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics* 23, 2021–2023.
- Mortensen, P., Gouw, J.W., Olsen, J.V., et al. (2010). MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J Proteome Res* 9, 393–403.
- Mueller, L.N., Brusniak, M.Y., Mani, D.R., and Aebersold, R. (2008). An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J Proteome Res* 7, 51–61.
- Mueller, L.N., Rinner, O., Schmidt, A., et al. (2007). SuperHirn—a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* 7, 3470–3480.
- Nesvizhskii, A.I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* 73, 2092–2123.
- Nesvizhskii, A.I., Keller, A., Kolker, E., and Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Analyt Chem* 75, 4646–4658.
- Nilse, L., Sturm, M., Trudgian, D., et al. (2010). SILACAnalyzer—A Tool for differential quantitation of stable isotope derived data. *Comput Intelligence Methods Bioinformatics Biostatistics* 6160, 45–55.
- Ong, S.E., Blagoev, B., Kratchmarova, I., et al. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 1, 376–386.
- Ong, S.E., and Mann, M. (2005). Mass spectrometry-based proteomics turns quantitative. *Nature Chem Biol* 1, 252–262.
- Orchard, S., Hermjakob, H., and Apweiler, R. (2003). The proteomics standards initiative. *Proteomics* 3, 1374–1376.
- Pan, C.L., Kora, G., McDonald, W.H., et al. (2006). ProRata: A quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. *Analyt Chem* 78, 7121–7131.
- Park, S.K., Venable, J.D., Xu, T., and Yates, J.R. (2008). A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat Methods* 5, 319–322.
- Pedrioli, P.G.A., Eng, J.K., Hubley, R., et al. (2004). A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 22, 1459–1466.
- Pratt, J.M., Simpson, D.M., Doherty, M.K., Rivers, J., Gaskell, S.J., and Beynon, R.J. (2006). Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat Protoc* 1, 1029–1043.

- Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Res* 12, 1231–1245.
- Reisinger, F., Krishna, R., Ghali, F., et al. (2012). jmzIdentML API: A Java interface to the mzIdentML standard for peptide and protein identification data. *Proteomics* 12, 790–794.
- Reiter, L., Rinner, O., Picotti, P., et al. (2011). mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat Meth* 8, 430–435.
- Schilling, B., Rardin, M.J., Maclean, B.X., et al. (2012). Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in Skyline: Application to protein acetylation and phosphorylation. *Mol Cell Proteomics* 11, 202–214.
- Schmidt, A., Kellermann, J., and Lottspeich, F. (2005). A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* 5, 4–15.
- Shadforth, I.P., Dunkley, T.P.J., Lilley, K.S., and Bessant, C. (2005). i-Tracker: For quantitative proteomics using iTRAQ I. *BMC Genomics* 6, 145.
- Shevchenko, A., Sunyaev, S., Loboda, A., et al. (2001). Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time of flight mass spectrometry and BLAST homology searching. *Analyt Chem* 73, 1917–1926.
- Shinkawa, T., Taoka, M., Yamauchi, Y., et al. (2005). STEM: A software tool for large-scale proteomic data analyses. *J Proteome Res* 4, 1826–1831.
- Silberring, J., and Ciborowski, P. (2010). Biomarker discovery and clinical proteomics. *Trac-Trend Anal Chem* 29, 128–140.
- Silva, J.C., Denny, R., Dorschel, C.A., et al. (2005). Quantitative proteomic analysis by accurate mass retention time pairs. *Analyt Chem* 77, 2187–2200.
- Tanner, S., Shu, H., Frank, A., et al. (2005). InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analyt Chem* 77, 4626–4639.
- Thiede, B., Hohenwarter, W., Krah, A., et al. (2005). Peptide mass fingerprinting. *Methods* 35, 237–247.
- Thompson, A., Schafer, J., Kuhn, K., et al. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analyt Chem* 75, 1895–1904.
- Tsou, C.C., Tsui, Y.H., Yian, Y.H., et al. (2009). MaXIC-Q Web: a fully automated web service using statistical and computational methods for protein quantitation based on stable isotope labeling and LC-MS. *Nucleic Acids Res* 37, W661–W669.
- Venable, J.D., Dong, M.Q., Wohlschlegel, J., Dillin, A., and Yates, J.R. (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* 1, 39–45.
- Wang, G.H., Wu, W.W., Pisitkun, T., Hoffert, J.D., Knepper, M.A., and Shen, R.F. (2006). Automated quantification tool for high-throughput proteomics using stable isotope labeling and LC-MSn. *Analyt Chem* 78, 5752–5761.
- Weston, A.D., and Hood, L. (2004). Systems biology, proteomics, and the future of health care: Toward predictive, preventative, and personalized medicine. *J Proteome Res* 3, 179–196.
- Wiese, S., Reidegeld, K.A., Meyer, H.E., and Warscheid, B. (2007). Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics* 7, 340–350.
- Xu, C., and Ma, B. (2006). Software for computational peptide identification from MS-MS data. *Drug Discov Today* 11, 595–600.
- Yates, J.R., 3rd, Park, S.K., Delahunty, C.M., et al. (2012). Toward objective evaluation of proteomic algorithms. *Nat Methods* 9, 455–456.
- Yu, C.Y., Tsui, Y.H., Yian, Y.H., Sung, T.Y., and Hsu, W.L. (2007). The Multi-Q web server for multiplexed protein quantitation. *Nucleic Acids Res* 35, W707–W712.
- Zhang, N., Aebersold, R., and Schwilkowski, B. (2002). ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* 2, 1406–1412.

Address correspondence to:
Faviel F. Gonzalez-Galarza
Institute of Integrative Biology
University of Liverpool
The Biosciences Building
Crown Street, L69 7ZB
Liverpool, U.K.

E-mail: F.Galarza@liv.ac.uk